Demonstrating Hybrid Provenance Queries from Script Runs

Yang Cao[1], Peter Slaughter[5], Christopher Jones[5], Matthew B. Jones[5], Qiwen Wang[2], Duc Vu[3], Priyaa Thavasimani[4], Qian Zhang[1], Timothy McPhillips[1], Paolo Missier[4], Lauren Walker[5], Dave Vieglais[6], Bertram Ludäscher[1 2]

[1]School of Information Sciences, University of Illinois at Urbana-Champaign
[2]Department of Computer Science, University of Illinois at Urbana-Champaign
[3]Department of Electrical and Computer Engineering, University of Illinois at Chicago
[4]School of Computing Science, Newcastle University, UK
[5]National Center for Ecological Analysis and Synthesis, UCSB, Santa Barbara, USA
[6] University of Kansas, Lawrence, USA

Data provenance is metadata that describes the origin and processing history of a data artifact. In the computational and data sciences, we can distinguish two forms of data provenance, i.e., *prospective* and *retrospective* provenance. The former describes the general *workflow* by which data is produced, while the latter consists of runtime observables, e.g., names, locations, and contents of files read and written during a script run (coarse-grained observables). Fine-grained provenance can be captured as well, e.g., changes to individual data records, program variables, parameter settings, etc. By *hybrid provenance,* we mean a combination of both provenance types, i.e., detailed retrospective provenance information is "situated" within the context of the conceptual-level workflow (prospective provenance), thus aiming to get the "best of both worlds": The appendix depicts such a hybrid graph with alternating process steps and data elements. The latter can be instantiated to document the specific files used in a concrete workflow run (see Appendix). Since computational and data science experiments can often last days, weeks, or even months and often require the execution of multiple scripts or workflows with varying input datasets and parameters, some of these script runs appear as chained together implicitly via intermediate data files, i.e., the output of one script run was subsequently used as an input to another script run. In this way, larger ad-hoc dataflow graphs are generated as part of a data-driven, exploratory workflow. By *multi-run provenance* we mean such dataflow graphs that link individual script runs via intermedia data files.

Our system demonstration (extending earlier work [1]; see also [2]) will illustrate the variety of provenance information that we are able to capture, query, and visualize using a combination of tools for exposing both prospective and retrospective provenance. We show how prospective provenance can be declared using YesWorkflow (YW) annotations that reveal the fine-grained (variable level) dataflow graph implicit in scripts, and how this prospective provenance can be integrated with the coarse-grained (file-level) retrospective provenance information recorded by the DataONE Run Managers for MATLAB and R. We demonstrate the usefulness of integrating prospective and retrospective provenance in this way with queries:

1. **Prospective provenance queries in the context of a single script.** This can expose and test data dependencies at the workflow-level.
2. **Retrospective provenance queries in the context of a single run of a single script:** captures actual input and output files of a script run and other runtime observables.
3. **Hybrid provenance query in the context of a single script and single run:** blends retrospective and prospective provenance, yielding new knowledge artefacts.
4. **Provenance query in the context of multiple scripts and multiple runs:** query and visualize data dependencies across multiple script runs

Our demonstration queries and provenance reports thus yield a more complete and comprehensible picture of data provenance from multiple script runs.

References

[1] Y Cao, D Vu, Q Wang, Q Zhang, P Thavasimani, T McPhillips, P Missier, B Ludäscher. YesWorkflow. Demonstration repository. https://github.com/idaks/dataone-ahm-2016-poster

[2] Q Zhang, Y Cao, Q Wang, D Vu, P Thavasimani, T McPhillips, P Missier, B Ludäscher. Revealing the Detailed Lineage of Script Outputs Using Hybrid Provenance. Submitted to IDCC 2017 (Practice Paper track).

Appendix 1: Ocean Health Index for Howe Sound, British Columbia Example

1. Prospective provenance graphs produced by YW (see Figure 1)
2. Hybrid provenance graphs produced by YW and DataONE RunManager R client (see Figure 2)

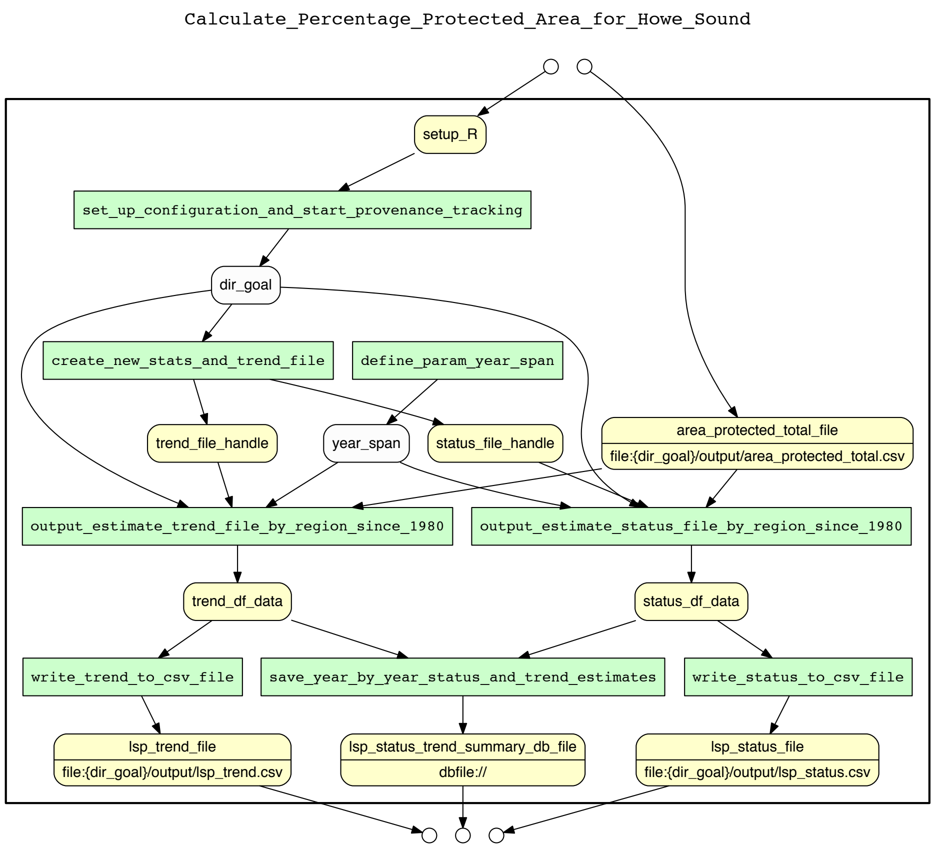
Calculate_Percentage_Protected_Area_for_Howe_Sound
setup_R
set_up_configuration_and_start_provenance_tracking
dir_goal
create_new_stats_and_trend_file
define_param_year_span
trend_file_handle
year_span
status_file_handle
area_protected_total_file
file:{dir_goal}/output/area_protected_total.csv
output_estimate_trend_file_by_region_since_1980
output_estimate_status_file_by_region_since_1980
trend_df_data
status_df_data
write_trend_to_csv_file
save_year_by_year_status_and_trend_estimates
write_status_to_csv_file
lsp_trend_file
file:{dir_goal}/output/lsp_trend.csv
lsp_status_trend_summary_db_file
dbfile://
lsp_status_file
file:{dir_goal}/output/lsp_status.csv

Figure 1 Complete YW prospective provenance graph for the `OHIBC_Howe_Sound` example. The diagram shows the workflow for calculating the "lasting special places sub goal" of the Ocean Health Index for Howe Sound, British Columbia that is based on the percentage of protected area within 3km of shore and 1km buffer zone inland, and how that has changed over time. The green box represents a computation step and a round yellow square box represents a data element declared in the script via YW tags.

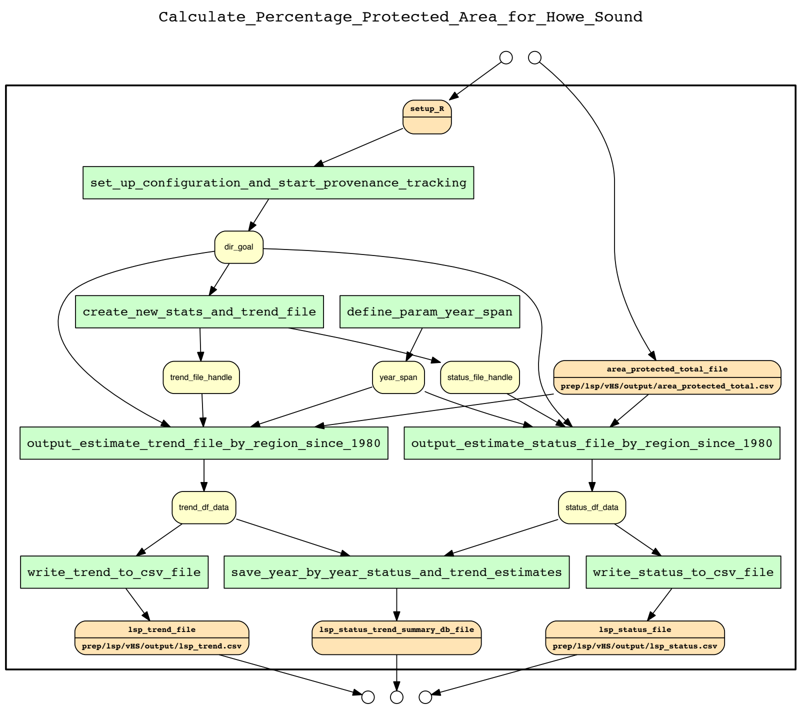
Calculate_Percentage_Protected_Area_for_Howe_Sound
setup_R
set_up_configuration_and_start_provenance_tracking
dir_goal
create_new_stats_and_trend_file
define_param_year_span
trend_file_handle
year_span
status_file_handle
area_protected_total_file
prep/lsp/vHS/output/area_protected_total.csv
output_estimate_trend_file_by_region_since_1980
output_estimate_status_file_by_region_since_1980
trend_df_data
status_df_data
write_trend_to_csv_file
save_year_by_year_status_and_trend_estimates
write_status_to_csv_file
lsp_trend_file
prep/lsp/vHS/output/lsp_trend.csv
lsp_status_trend_summary_db_file
lsp_status_file
prep/lsp/vHS/output/lsp_status.csv

Figure2 Complete reconstructed hybrid provenance graph for the OHIBC_Howe_Sound example using YW and DataONE RunManager R client. In the diagram, the runtime file-level observables are augumented with YW prospective graph.

Appendix 2: Carbon 3/Carbon 4 (C3/C4) soil maps for North America example

1. Hybrid provenance graph produced by YW and RunManager MATLAB client (see Figure 3)
2. Retrospective provenance query in the context of multiple scripts and multiple runs (see Figure 4)

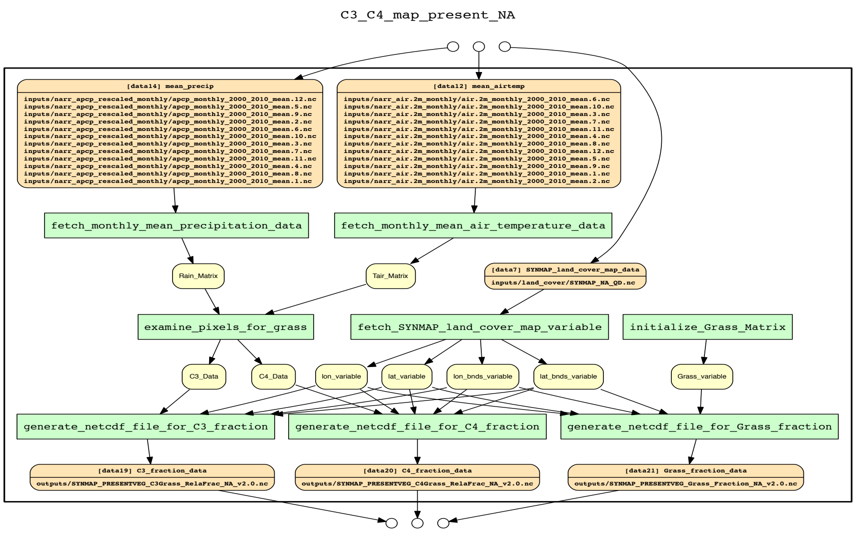
C3_C4_map_present_NA
[data14] mean_precip
inputs/narr_apcp_rescaled_monthly/apcp_monthly_2000_2010_mean.12.nc
inputs/narr_apcp_rescaled_monthly/apcp_monthly_2000_2010_mean.5.nc
inputs/narr_apcp_rescaled_monthly/apcp_monthly_2000_2010_mean.9.nc
inputs/narr_apcp_rescaled_monthly/apcp_monthly_2000_2010_mean.2.nc
inputs/narr_apcp_rescaled_monthly/apcp_monthly_2000_2010_mean.6.nc
inputs/narr_apcp_rescaled_monthly/apcp_monthly_2000_2010_mean.10.nc
inputs/narr_apcp_rescaled_monthly/apcp_monthly_2000_2010_mean.3.nc
inputs/narr_apcp_rescaled_monthly/apcp_monthly_2000_2010_mean.7.nc
inputs/narr_apcp_rescaled_monthly/apcp_monthly_2000_2010_mean.11.nc
inputs/narr_apcp_rescaled_monthly/apcp_monthly_2000_2010_mean.4.nc
inputs/narr_apcp_rescaled_monthly/apcp_monthly_2000_2010_mean.8.nc
inputs/narr_apcp_rescaled_monthly/apcp_monthly_2000_2010_mean.1.nc
[data12] mean_airtemp
inputs/narr_air.2m_monthly/air.2m_monthly_2000_2010_mean.6.nc
inputs/narr_air.2m_monthly/air.2m_monthly_2000_2010_mean.10.nc
inputs/narr_air.2m_monthly/air.2m_monthly_2000_2010_mean.3.nc
inputs/narr_air.2m_monthly/air.2m_monthly_2000_2010_mean.7.nc
inputs/narr_air.2m_monthly/air.2m_monthly_2000_2010_mean.11.nc
inputs/narr_air.2m_monthly/air.2m_monthly_2000_2010_mean.4.nc
inputs/narr_air.2m_monthly/air.2m_monthly_2000_2010_mean.8.nc
inputs/narr_air.2m_monthly/air.2m_monthly_2000_2010_mean.12.nc
inputs/narr_air.2m_monthly/air.2m_monthly_2000_2010_mean.5.nc
inputs/narr_air.2m_monthly/air.2m_monthly_2000_2010_mean.9.nc
inputs/narr_air.2m_monthly/air.2m_monthly_2000_2010_mean.1.nc
inputs/narr_air.2m_monthly/air.2m_monthly_2000_2010_mean.2.nc
fetch_monthly_mean_precipitation_data
fetch_monthly_mean_air_temperature_data
Rain_Matrix
Tair_Matrix
[data7] SYNMAP_land_cover_map_data
inputs/land_cover/SYNMAP_NA_QD.nc
examine_pixels_for_grass
fetch_SYNMAP_land_cover_map_variable
initialize_Grass_Matrix
C3_Data
C4_Data
lon_variable
lat_variable
lon_bnds_variable
lat_bnds_variable
Grass_variable
generate_netcdf_file_for_C3_fraction
generate_netcdf_file_for_C4_fraction
generate_netcdf_file_for_Grass_fraction
[data19] C3_fraction_data
outputs/SYNMAP_PRESENTVEG_C3Grass_RelaFrac_NA_v2.0.nc
[data20] C4_fraction_data
outputs/SYNMAP_PRESENTVEG_C4Grass_RelaFrac_NA_v2.0.nc
[data21] Grass_fraction_data
outputs/SYNMAP_PRESENTVEG_Grass_Fraction_NA_v2.0.nc

Figure 3 Complete reconstructed hybrid provenance graph for the C3C4 example (version 1) using YW and DataONE RunManager MATLAB client. The example script (in MATLAB) produces Carbon 3/Carbon 4 (C3/C4) soil maps for North America using average rain and air temperature monthly data from year 2000 to 2010.

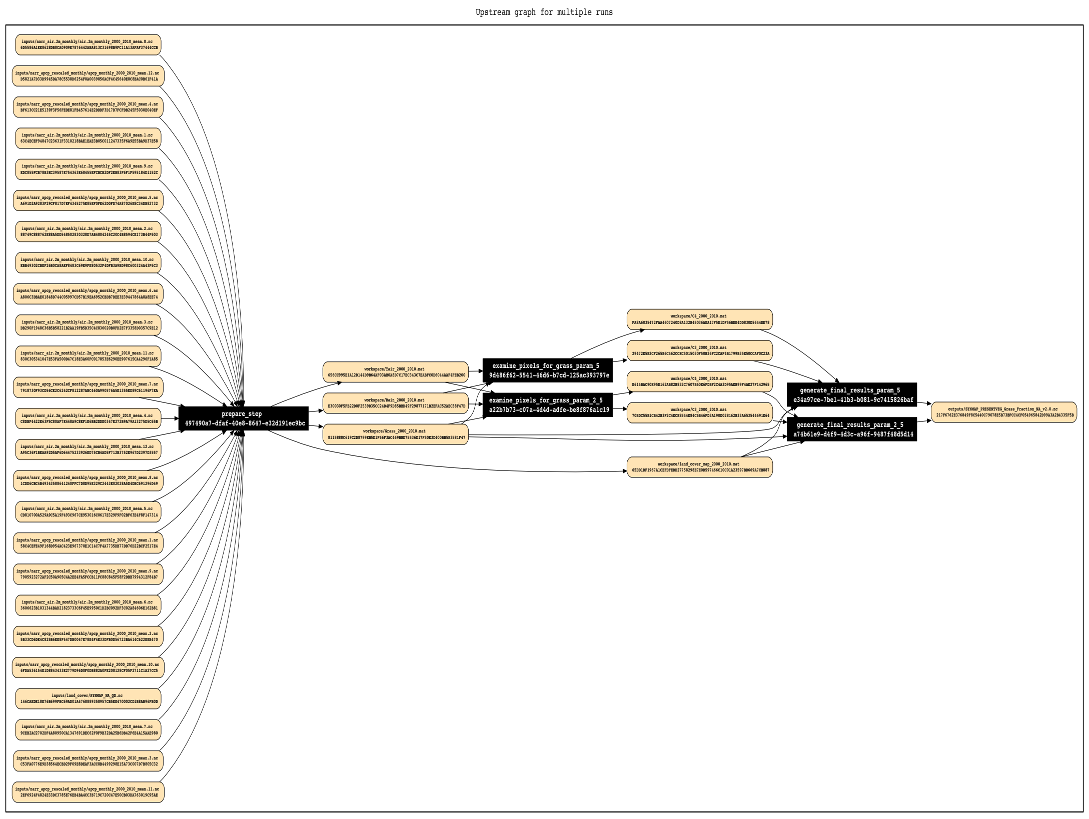
Upstream graph for multiple runs
prepare_step
497490a7-dfaf-40e8-8647-e32d191ec9bc
workspace/Tair_2000_2010.mat
656CC995E1A12D144D9B64AF03AB0A8D7C17EC343C7EABFC0D6044AAF4FEB200
workspace/Rain_2000_2010.mat
E30030F5FB22D0F2539D35CC24D4F9D85BBD49F29877171B2EFAC52AEC38F47D
workspace/Grass_2000_2010.mat
8115888C619C2D8799EB5D1F66F3AC469BBD75536D17F50E3D400BB5E3581F47
examine_pixels_for_grass_param_5
9d486f62-5541-46d6-b7cd-125ac393797e
examine_pixels_for_grass_param_2_5
a22b7b73-c07a-4d4d-adfe-be8f876a1c19
workspace/C4_2000_2010.mat
FA8A6035472FAA4607240D8A132B45036AEA17F5D1DF56BDE4DD830D5644DD78
workspace/C3_2000_2010.mat
29472E5B2CF265B6C462CCEC5015030F50B26FC2CAF4B1799B35E50CCAF0C23A
workspace/C4_2000_2010.mat
E614BAC90E95D162AB82B832C7607B60E6FDBF2C4A3D95AEB99F4AE27F142965
workspace/C3_2000_2010.mat
70BDC55B1CB42B3F2C4ECE8544E84C6B46FD3A190D028162B33A653544691E64
workspace/land_cover_map_2000_2010.mat
65DD1DF1967A1CEFDFEDD27758298E7E0D597466C10C01A23597BD669A7CB887
generate_final_results_param_5
e34a97ce-7be1-41b3-b081-9c7415826baf
generate_final_results_param_2_5
a74b61e9-d4f9-4d3c-a96f-9487f48d5d14
outputs/SYNMAP_PRESENTVEG_Grass_Fraction_NA_v2.0.nc
217F6762E376849F8C5440C79078E5B73BFCC4CF054965842D99A3A2B4335F5B

Figure 4 Upstream retrospective provenance subgraph (longitudinal view) recording multiple scripts and multiple runs for `C3C4` example. The generated graph recorded a set of runs that consist of three scripts, five runs and the second script was run with two different parameter values (i.e., 2.5 and 5) twice. The example script set is called `C3C4` example version 2. In the diagram, an orange yellow box represents a data file and a black box represents an execution (a run). When two data files sharing with a single orange yellow box, it indicates that these two files have the identical contents.